International Core Journal of Engineering Volume 8 Issue 3, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0061

Rumour Detection and Analysis on Twitter based on COVID-19

Lihua Wang

School of Computing and Information System, The University of Melbourne, Parkville, VIC 3010, Australia

Abstract

Rumour detection on Twitter is a significant issue due to its economic and social impact. Nowadays, the problem of rumour detection on twitter dataset addressed by applying a number of machine learning and deep learning models, including SVM, C-LSTM and BERT, to perform rumour analysis on source tweets as well as reply tweets. In this paper, the model of trained BERT rumour classifier from the first task applied to a set of provided COVID-19 tweets dataset to detect rumours and analyze how topics, users and emotions of rumours differ from non-rumours.

Keywords

Rumour Detection; Machine Learning; COVID-19.

1. Introduction

Generally, rumours spread rapidly via Twitter and can have a huge economic and social impact. In this manuscript, the rumour detection on "tweets" are conducted by various machine learning algorithms. Commonly, two possible ways for the rumour detection are investigated: Linguistic Approach and Network Approach. For the former, the rumour content should be extracted based on language patterns. The misinformation would have some similar language patterns such as fear mongering. While for the latter, the metadata for the information can be taken into consideration, such as the user characters in this case. A tweet from an unreliable source (rumour spreader user) is likely to be a rumour, thus, the Linguistic Approach as described are considered and identify rumours based on only text information (tweets "text").

2. Methodology

2.1 Dataset Pre-processing

Twitter data have to be normalized to create a dataset that can be easily learned by classifiers since tweets have certain special characteristics such as emotions, user mentions, URLs, etc. There are several pre-processing steps performed to clean the raw data and reduce its size as the following seven section:

1) Using the NLTK Tweet Tokenizer package to tokenize the tweets;

2) Removing all the URLs by matching expression (www\.[\S]+)|(https?://[\S]+);

3) Removing all the user mentions by matching expression (@[^\s]+);

4) Converting the tweet to lower case;

5) Removing single character words, non-English words and other words;

6) Custom stopwards expanded from NLTK stopword list;

7) Lemmatizing all words by applying WordNetLemmatizer package tweet.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0061

2.2 Feature Engineering

Most rumour spreaders would use certain language strategies to increase its credibility to the audience. Therefore, how the text features can capture the language patterns for rumours will directly impact the accuracy of the whole rumour detection system. In this paper, three type of word features representation:

Most rumor mongers will use some language strategies to improve their credibility to the audience. Therefore, how text features capture the language pattern of rumors will directly determine the accuracy of the whole rumor detection system. Herein, in this paper, three typical word features, including BOW, TF-IDF and word sequence in different models are investigated as followed:

BOW: BOW believes that the post-processed words in a tweet text as separated by equally important makers. This representation method can provide a lot of information about the text itself. Just as in our problem the words in rumours much likely to express an unobjective emotion or negative situation.

TF-IDF: TF-IDF in SVM models was used in this paper, however, this method can only represent simple word mentions in texts and unable to capture syntax or styles of rumours. Rumours and non-rumours may share the similarity of tokens, but the different language syntax and structure will lead to completely different meaning.

Word Sequence: Sequence respecting approaches have an edge over BOW when a synthetic tweet corpus dominated by sentences and consideration to the position of the words are important. When some neural network models were trained, some deep learning approaches such as LSTM, GNU are applied to model the tweets as embedding string of words.

2.3 Models

Typically, three models are used in the project to create the detection system. One is baseline model using binary class classification SVM. Other two for model improvements including neural network mix model CLSTM, and BERT. “train.json” dataset will be used in training process and hyper parameter tunning and evaluation will be conducted on “dev.json” dataset.

Note: In both train and dev datasets are divided into source tweets and source + reply tweets, training and evaluating respectively.

SVM: SVM is a simple non-probabilistic binary linear classifier. In order to compare the data feature performance between bow and TF-IDF, the post-processing twitter data model is constructed based on these two features. At the same time, the experiment is also realized through the SVC classification in the scikit learning SVM package in Python, where C = 1.0, using the fine-tuning radial basis function kernel.

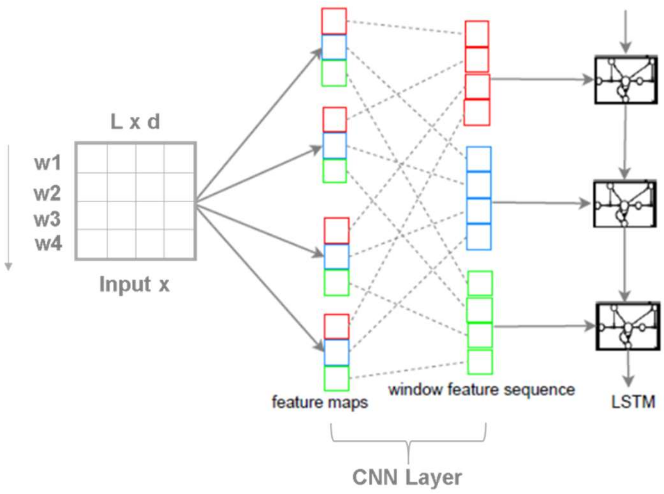

Figure 1. C-LSTM model description [5]

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0061

C-LSTM: keras was used with TensorFlow backend to implement the C-LSTM Neural Network model, which combined the strengths of RNN and CNN architectures for sentence representation and text classification. C-LSTM utilizes CNN to extract a sequence of higher-level phrase from source tweets, then the vectors generated from the CNN are fed into a LSTM recurrent neural network for rumour detection for tweets. The details are illustrated in Figure 1. In the implementation of this manuscript, after initializing the first layer embeddings of sequential inputs, the output from the CNN layer (second layer) was directly concatenated with LSTM layers (third layer).

Fine-Tuning BERT: Thinking about the tweets data between the source and reply will have a contextual relationship, as well as a time-based relationship after sorted by tweets created time, I used the BERT BASE since it is a bi-direction trained transformer language model which contains an attention mechanism to learn the contextual relationship between words and then generate deep bidirectional context representations by jointly conditioning on both left and right context in all layers. In the implementation (Figure 2), after feeding the input sequence to BERT and pass the contextualised embedding of [CLS] produced by BERT to a simple feedforward network classifier, one additional output layer with a SoftMax function for rumour classification was added. Thus it can produce a single scalar value to denote the probability of the rumour class. Also, fine-tuning in BERT that extends the training max length to 256 was applied, since the average input sentences distributed between 150 to 200.

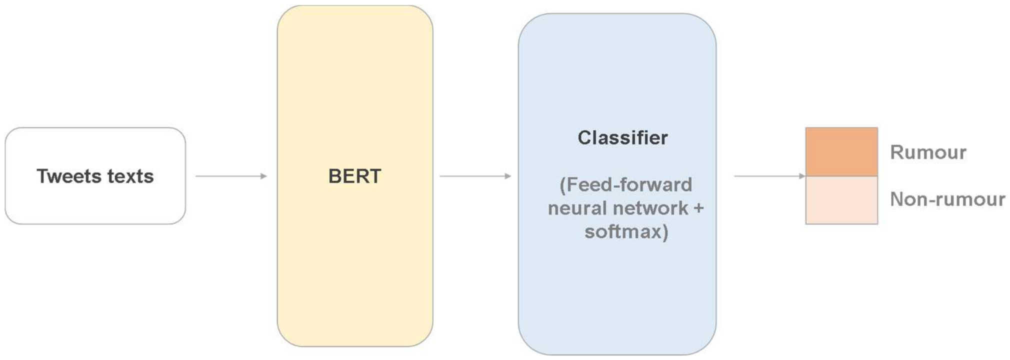

Figure 2. BERT model description

3. Evaluation and Analysis

SVM: Based on the results (Table 1), SVM performs badly and fails due to the very low recall across all classes and results in low F1 scores (Average69%), however, it performs well in precision above 0.8, indicating that the model classifies a few negative cases as rumour, it shows that SVC classifier tends to predict a label as positive. It is can be seen that the TF-IDF representation recorded the best performance in source tweets since TF-IDF will filter out irrelevant (noise) words and with training only on the source tweets. Additionally, it has less dimensions to expend which results in higher efficient since SVM has a poor explanatory power for high-dimensional mappings of kernel functions, especially radial basis functions.

C-LASTM: CNN and RNN, C-LSTM models are listed to compare the performances. According to the results (Table `), they all have a significant improvement on F1 score (Average 78%) compared with SVM, though it still not up to expectation. C-LSTM performs best on F1 score, where CNN had a slightly lower score following the worst performance on RNN. This can be explained by that sequential processing over time works well on whole tweets, with combined the advantages of features extraction from CNN and long context memory from LSTM. However, the model performance is unstable, and computation and time consumptions are not efficient.

BERT: BERT performs consistently excellent for each class (Table 1) and reach to the highest F1 score at 81.73% on the whole dataset (source + reply tweets). That means, BERT can adjust the weights associated with the model to better represent text. In more specific, during classifier fine-tuning, the starting points of the weights are closer to values that correctly model Twitter data. However, there is a huge time consumption during the training process because of the slow converges on model since only 15% of the data in each batch size is involved in the prediction.

Table 1. Result of Models

Model	Inputs	F1	R	P
SVM	TF-IDF (Source Tweets)	0.69	0.66	0.81
	TF-IDF (Whole Tweets)	0.67	0.66	0.79
	BOW (Source Tweets)	0.67	0.69	0.8
	BOW (Whole Tweets)	0.64	0.64	0.76
RNN	Source Tweets	0.75	0.67	0.84
CNN	Source Tweets	0.77	0.80	0.81
C-LSTM	Source Tweets	0.75	0.82	0.74
	Whole Tweets	0.78	0.86	0.76
BERT	Source Tweets	0.77	0.81	0.84
	Whole Tweets	0.80	0.80	0.87

Comparison Between Three Models: BERT has an obviously higher F1 score and more stable performance than the others. It proves again that a Bi-direction trained language model has a deeper understanding of the context than a one-way language model, which becomes to the most suitable model to process the tweets dataset.

4. COdaLab Competition and Conclusion

Finally, an ensemble method was used to take a majority vote over the predictions of the best performance in the above 3 different models. The result shows an accuracy of 83.39%.

5. Tweets Analysis about Covid-19 based Issues

5.1 About Covid-19 in Tweets

Since the beginning of 2020, the COVID-19 pandemic has come to dominate both traditional news and social media platforms, and misinformation such as fake news, conspiracy theories and rumours thrive during these uncertain times. The aim of this section is to figure out what kinds of COVID-19 rumours are being distributed on Twitter. After applying the best performing rumour detection model from part 1 to the given Covid-19 based tweets dataset, projected labelled rumour data could be obtained to distinguish rumours from non-rumours in this COVID-19 data and used to analyse the users' characteristics and tweets texts. In this dataset, the findings show a variety of fascinating observations about users, topics, and emotions. For example, rumour-spreaders, usually have low follower but a high following count and they prefer to discuss politics (mostly party blaming), are more emotionally charged (e.g., anger) with more negative sentiments.

5.2 Data Design

1548 events were labelled as “rumour” and 15910 events were labelled as “non-rumour”, after applying the BERT rumour detection on the total 17458 events in COVID-19 dataset in Tweets, By counting the retweet and tweet counts on each day, we can visualize the volume of non-rumour and rumour tweets over time, respectively. We can see that they both have some traffic of COVID-19 related tweets from late January 2020, although they do not really pick up until early-March and I suspect the spike of activity may be due the World Health Organisation declaring it as a pandemic on 12th March 2020. There is a significant high point on the rumour figure at around the date of 24 April 2020, I supposed this may be due the rumour words from Trump that the COVID-19 can be treated by injecting disinfectant which has drawn much attention to discuss.

In terms of pre-processing, the NLTK Tweet Tokenizer package was conducted to tokenize the tweets, and lowercase and lemmatise all words by applying WordNetLemmatizer package, as well as eliminate digits, hyperlinks and @usernames. I also use an expanded NLTK stopword list to filter stopwords, which includes COVID-19-related stopwords like covid19 and coronavirus. For topic analysis, hashtags are also omitted.

5.3 Results and Analysis

Topic analysis: To find the popular topics discussed in datasets, the top frequent unigram and bi-gram words that extracted from source tweets (Herein, I did not add reply tweets since in some case, reply tweet might interference with the central topics of a tweet event.) was retrieved and the frequency dictionaries in Word cloud format was visualized as well as line charts (Figure 3). After comparison, bigram word clouds show more reasonable results and we can see several broad topics: (1) COVID19-status reports (tested positive, confirmed case, new case, report new, death toll); (2) health advice (social distancing, public health, and wear mask) appeared frequently in both two classified datasets where the topic of US politics (president trump and white house) accounts for the largest proportion for non-rumours, and COVID19-status reports for rumours. To better understand the topical difference between the two datasets over time changes, I tracked both bigram dictionaries sorted by moths from January to July and Table 2 gives an overview of the top frequent topics over a month. As the pandemic grew, there is no surprise that the emerged keywords of rumours associated starting with "Wuhan epidemic" to "new cases", "death tolls" and "pandemic began" in the end. Looking at non-rumours, the topics are very different: they are mostly related to "president trump" and health advice and turn to focusing on covid cases status in the late period. The reason I supposed that Trump's highly active Twitter rumour remarks aroused heated discussion and dissatisfaction among people.

Hashtag analysis: To find the most popular hashtags of COVID-19 rumours and non-rumours, this study separately extracted the hashtag words (start from the first character "#") from the rumour and non-rumour raw source and reply tweets and filtered COVID-19-related hashtags like "#covid19". After sorting the filtered hashtags frequencies and representing the top 20 popular hashtags of non-rumours and rumours as bar charts (Figure 3h and 4). An Interesting finding is that the non-rumours figure shows a similarity with the above popular topics as the president trump still be a highly topical theme associated with many sarcasm and accusation emotions hashtags like "#TrumpVirus", "#TrumpLiesPeopleDie". For rumours, most popular hashtags related to the healthcare such as personal protection equipment and, unsurprisingly, "#WuhanVirus" and "#China" are also labelled as rumour hashtags.

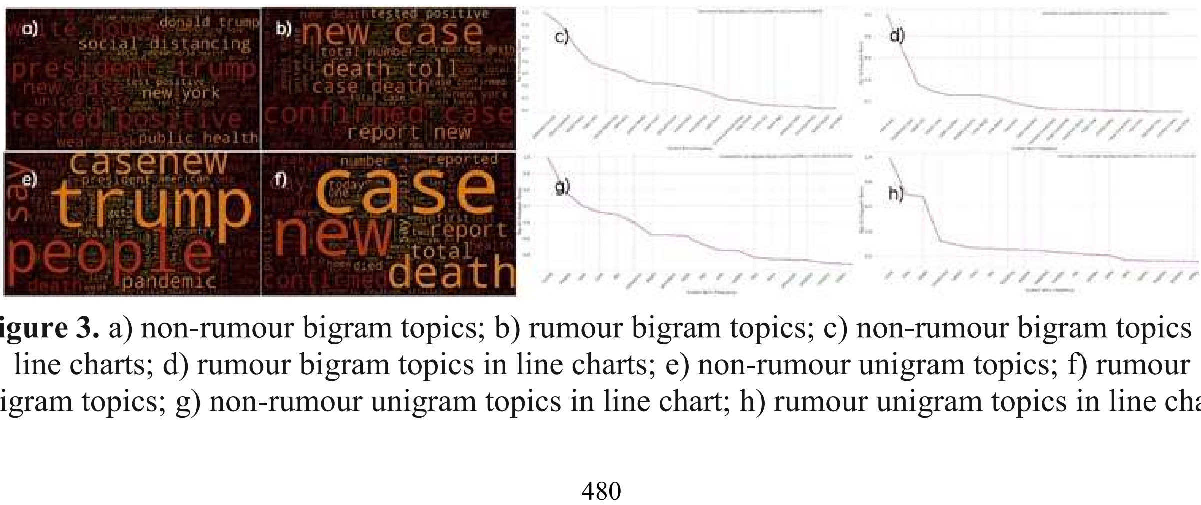

Figure 3. a) non-rumour bigram topics; b) rumour bigram topics; c) non-rumour bigram topics in line charts; d) rumour bigram topics in line charts; e) non-rumour unigram topics; f) rumour unigram topics; g) non-rumour unigram topics in line chart; h) rumour unigram topics in line chart.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0061

Table 2. Trending bigrams on rumours and non-rumours over months flow

Month	non-rumours	rumours
01	case reported, yes frightening, year celebration	confirmed case, wuhan epicentre, went effect
02	trump administration, public health, washing hand	get flu, flu shot, year see, went doctor, washington say
03	president trump, tested positive, stay home, social distancing	confirmed case, new case, report new, total case, new york, Italy report
04	president trump, white house, tested positive, social distancing, stay home	death toll, new case, confirmed case, total number, case reported, case death, new york
05	president trump, tested positive, nursing home, new case, trump say, wear mask	new case, confirmed case, death toll, tested positive, nursing home
06	new case, tested positive, wear mask, president trump	new case, confirmed case, report new, single day, juen juen
07	tested positive, wear mask, president trump, new case	new case, pandemic began, case today, new death

User analysis: This study only on users who published the source tweets in analysis. Table 3 presents some statistics of these users for rumours and non-rumours and Figure 5 visualizes the comparison of some items of rumour and non-rumour users' characters. Interestingly, users who are involved in rumour creation tend to tweet more (higher post counts) rather than have fewer posts that are tagged as a favourite, compared with non-rumour users.

Table 3. User Statistics

	non-rumours	rumours
#Folower	5409719	4776971
#Following	7931	7196
Ratio (#Follower / #Following)	682	663
#Faverate Counts	14299	5500
#Post	112456	151732
Account Created Date	2011-04-30	2010-12-01
Geo Enabled	42%	40%

Besides, the rumour users have slightly fewer followers as well as the followings. Their account is also generally older than non-rumour users that the average created date is on 2010-12-01, whereas the non-rumours user accounts created at about 2011-04-30. I also retrieve the most mentioned users in raw tweets and selected top 20 @Username in rumours and non-rumours, respectively (Figure 6). Except the @RealDonaldTump being the most popular mentioned user in both classifications, it is seen that some famous rumour news disseminators, such as @ foxbusiness and @ CNN, are marked as rumour accounts. On the other hand, @ joebiden has become an obvious non rumour user. Since compared with trump's unreliable policies in the epidemic, Biden's administration has received more public support.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0061

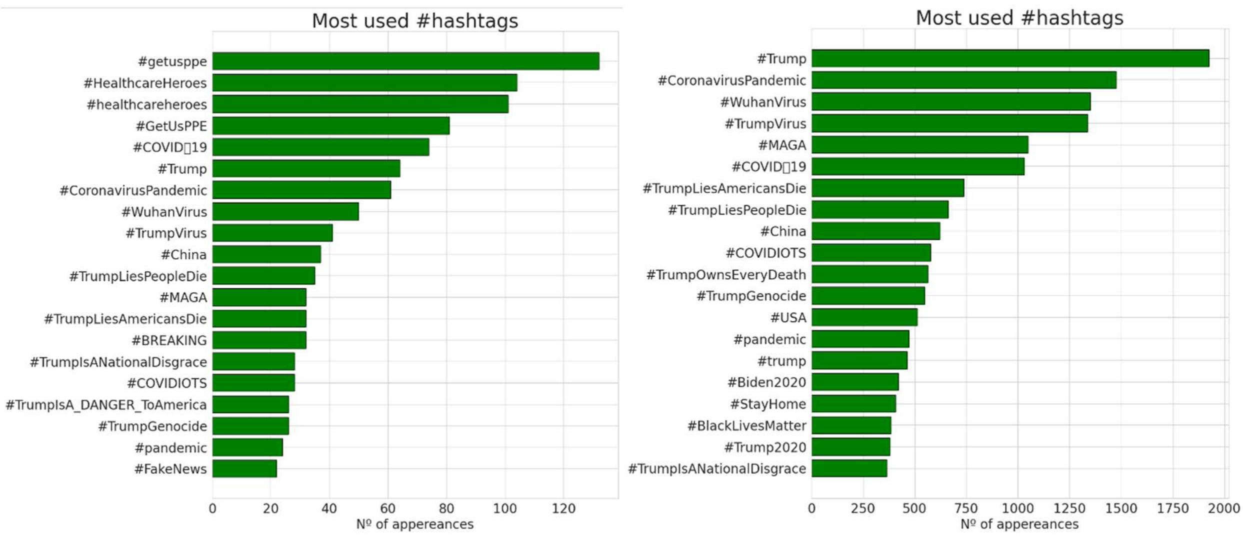

Figure 4. a) non-rumours top 20 popular hashtags; b) rumours top 20 popular hashtags.

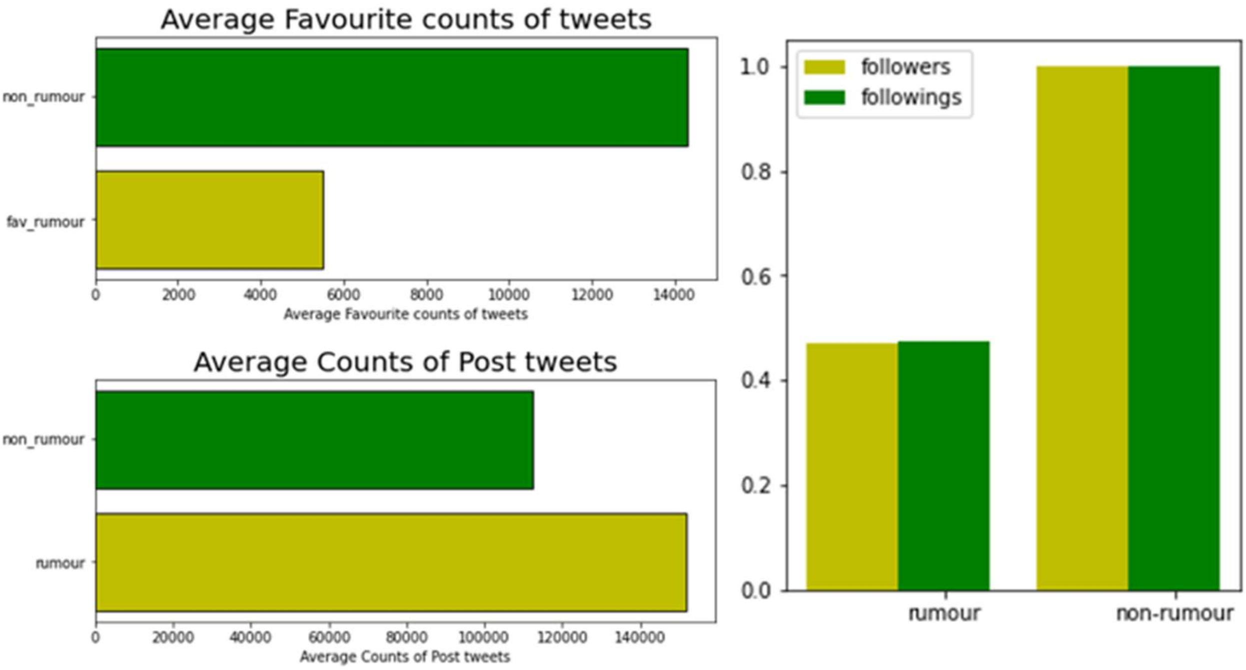

Figure 5. user statistic visualization

Sentiment/Emotion Analysis: An emotion prediction system to classify the emotion of tweets in the data to better recognize the public's sentiments during the COVID19 crisis. DeepMoji [1], a Bi-LSTM with an attention model trained on a large number of emoji occurrences in tweets, is used throughout this experiment and I labelled the source tweets data with 63 predefined emojis by applying their pre-trained model. Figure 7 illustrates the distribution of emojis for source tweets in rumours and non-rumours. Compared the two-emotion pie charts, we see a similar distribution for the top 5 emotions (👍, 😠, 😒, 👏, 💪), but a bit confused observation here is the top emoji for rumours is 👍 account for 9.07% whereas 😒 as the top emoji for non-rumours accounts for 8.88%, though the percent difference is less severe. Overall, based on the emoji represents, the emotions of the public generally express the anger and disagreement, the rest of them usually express encouragement, as well as a pray from the source tweets observation.

Volume 8 Issue 3, 2022
DOI: 10.6919/ICJE.202203_8(3).0061

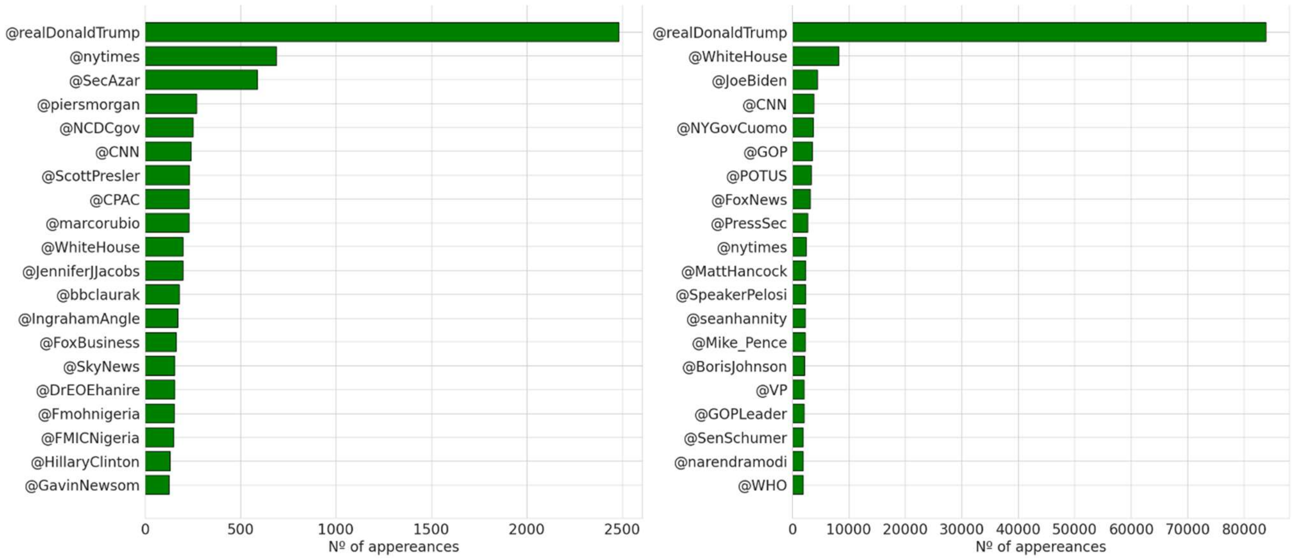

Figure 6. a) most active @users in non-rumours; b) most active @users inrumours.

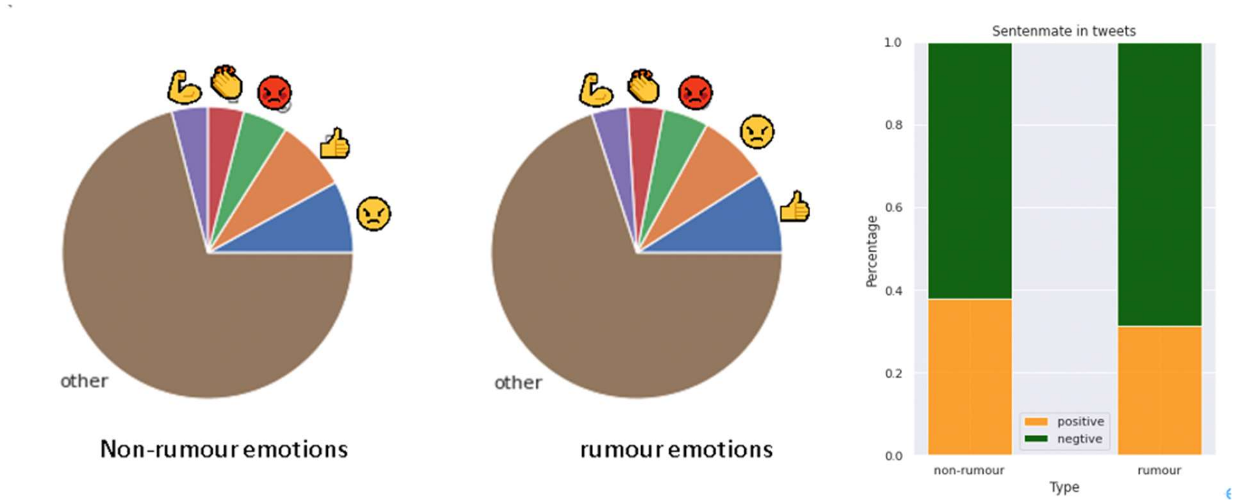

Figure 7. non-rumour emoji emotions, rumour emoji emotions, and sentiments (From left to right).

To accurate the sentiments of the public from tweets, another sentiment detected model, package Baidu Senta Corpus, was applied.[2]. A type of sentiment Analysis system to interpretate and classify the sentiments (positive and negative) within text data. Figure 7 reveals the public sentiments towards source tweets and we can see that negative sentiment dominates both rumours and non- rumours, but substantially more in rumours than non-rumours (68% vs. 62%).

6. Conclusion

In conclusion, as a hot research topic, moving target tracking technology has been widely investigated in various fields. Also, as mentioned above, with the help of low cost, low power consumption, self-organization and high error tolerance of wireless sensor networks, moving target tracking based on wireless sensor networks also has broad application prospects. In short, after implementing several ML algorithms: SVM, C-LSTM, BERT to build the rumour detection system, It is found that the performance of Bert on twitter data set is better than other models, which proves that bidirectional transformation is more suitable for dealing with Twitter data. Additionally, the F1 score of 83.98% on CodaLab leader board was arrived via majority voting from 3 best predictions of the above 3 models. Finally, the BERT model to classify COVID-19-related rumours on given tweet datasets,

ISSN: 2414-1895 DOI: 10.6919/ICJE.202203_8(3).0061

provided a quantitative test to demonstrate analysis of rumours vs. non-rumours users, topics, and emotions, and discovered a number of insights.

References

[1] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, S. Lehmann, Using millions of emoji occur- rences to learn any-domain representations for detecting sentiment, emotion and sarcasm, in: Proceedings of the 2017 Conference on Empir- ical Methods in Natural Language Processing, Copenhagen, Denmark, 2017, pp. 1615–1625.

[2] Tian, H., Gao, C., Xiao, X., Liu, H., He, B., Wu, H., ... & Wu, F. (2020). SKEP: Sentiment knowledge enhanced pre-training for sentiment analysis. arXiv preprint arXiv:2005.05635.

[3] Tian, L., Zhang, X., & Lau, J. H. (2020). # Democrats are destroying America: Rumour analysis on twitter during COVID-19. In CEUR Workshop Proc..Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. 1981. Alternation. Journal of the https://doi.org/10.1145/322234.32224.

[4] Tian, L., Zhang, X., Wang, Y., & Liu, H. (2020, April). Early detection of rumours on twitter via stance transfer learning. In European Conference on Information Retrieval (pp. 575-588). Springer, Cham.

[5] Zhou, C., Sun, C., Liu, Z., & Lau, F. (2015). A C-LSTM neural network for text classification. arXiv preprint arXiv:1511.08630.